COMPARISON OF OBSERVATIONAL RECORDINGS IN VIVO, VIA MIRROR, AND VIA TELEVISION

RONALD N. KENT, K. DANIEL O'LEARY, ALLEN DIETZ,
AND CHARLES DIAMENT

STATE UNIVERSITY OF NEW YORK AT STONY BROOK

Simultaneous observational recordings were made *in vivo*, *via* an observation mirror, and *via* closed circuit television. Three of nine observers had extensive experience recording behavior *in vivo*; three had extensive experience recording behavior *via* mirror; and three had extensive experience recording *via* television. Observers recorded nine categories of disruptive behavior for children in a special class setting. Frequencies of behavior recorded *in vivo*, *via* mirror, and *via* television differed significantly for only one category, vocalization. There were no significant main effects or interactions involving the observers' previous experience. Occurrence reliability coefficients computed within and between media demonstrated the similarity of observer agreement in all three media. Data collection procedures using an observational mirror or closed circuit television appear to be reasonable alternatives to *in vivo* observation in circumstances similar to those in the present study.

DESCRIPTORS: observation, *in vivo*, tv, mirror, reliability

Observational recording has long been utilized as a major data source in developmental, social, and clinical psychology. However, recently a variety of concerns have been expressed regarding possible limitations of observational procedures which are particularly relevant when recording is executed *in vivo*. Alteration of behaviors of interest as a function of observer presence (*cf.* Mercatoris and Craighead, 1974), biases in observer judgment as a function of salient experimental hypotheses or procedures (*cf.* Kent, Diament, Dietz, and O'Leary, 1974; O'Leary, Kent, and Kanowitz, 1975), and inconsistency over time and between observers (*cf.* O'Leary and Kent, 1973) have been found to be possible sources of inaccuracy in the recording of behavior associated with *in vivo* recording procedures.

This research was supported by Grant MH 21813 from the National Institute of Mental Health and Grant OEG-0-71-2872-607 from the Office of Education. Requests for reprints may be sent to K. Daniel O'Leary, Department of Psychology, State University of New York at Stony Brook, Long Island, New York 11794. Ronald Kent is now at Brown University, Providence, Rhode Island. Allen Dietz is now at the University of Texas, Austin, Texas. Charles Diament is now at Children's Psychiatric Center, Eatontown, New Jersey.

The use of observation mirrors or videotape cameras may reduce or eliminate some of these potential inaccuracies. Clearly, shielding the presence and reactions of observers, if not the actual fact of observation, from their subjects may decrease reactivity in field-experimental settings. Videotapes provide additional advantages. They may be edited to conceal the introduction and elimination of treatment procedures, thereby reducing possible biases due to expectations of the observer. Further, the order in which videotapes are viewed by observers may be varied to eliminate the application of differential recording criteria to sequential treatment phases.

The purpose of the present study was to compare systematically the reliabilities and frequencies of simultaneous observational recordings of children in a classroom setting obtained *in vivo*, through a one-way mirror, and *via* closed-circuit television.

METHOD

Observers and Experimenters

Nine undergraduates with one semester of experience using an observational code for dis-

ruptive behavior in a classroom setting served as observers in partial fulfillment of a course requirement. Three of these observers had been trained and had observed solely *in vivo* over the course of the previous semester, three through a one-way mirror, and three *via* a television monitor. The two experimenters were an undergraduate student and a bachelor's level research assistant, each of whom had over 2 years experience in using the observational code.

Setting

All training and experimental sessions involved observation of a special class for first- and second-grade children with academic and emotional problems at Point of Woods Laboratory School, State University of New York at Stony Brook. Detailed rating scale and achievement data of the children as well as a detailed curriculum description are provided in S. O'Leary and Schneider (1977). This classroom is adjacent to an observation room equipped with a one-way mirror (1.2 m $\times$ 2.4 m), an audio amplification system (Lafayette LA-25, 20 watt), and a videotape camera (SONY, AV-3200). Two microphones (SONY, ECM 19B) were located in opposite corners of the classroom.

Behavioral Code

The code for disruptive behavior in a structured classroom setting has been utilized in a variety of studies by O'Leary and his associates (*e.g.*, Drabman, Spitalnik, and O'Leary, 1973; Kaufman and O'Leary, 1972; O'Leary, Drabman, and Kass, 1973; O'Leary, Kaufman, Kass, and Drabman, 1970). The nine categories of this code were: out of chair, modified out of chair, touching, noise, vocalization, orienting, playing, aggression, and time-off-task. Observation involved noting the occurrence of each category of behavior during 20-sec intervals, which were separated by 10-sec intervals for recording. In addition, a composite score representing a sum of the nine individual behavioral categories per 20-sec interval was computed.

Training

The purpose of training sessions was to familiarize each of the observers with behavioral recording *in vivo* and *via* mirror and television (SONY, CVM, 192U). During these 1-hour sessions, observers were divided into three groups, each group of three composed of one person with experience observing *in vivo*, one *via* mirror, and one *via* television. Each group observed *via* each medium (*i.e.*, *in vivo*, mirror, and closed circuit television), computed reliability, and discussed disagreements with the others in their group. At no time during training were observers given the opportunity to compare observations with others who had observed simultaneously in a different medium. New groups were formed during each training session to reduce the development of between-group differences in rating. These groups were formed by systematically varying membership, with each of the members having had experience the previous semester in a medium different from the other members. Training for all observers was continued for 1 hour across 7 days, at which time each observer maintained an average occurrence reliability across behaviors of over .70 for at least 2 consecutive days.

Experimental Sessions

The experimental design crossed the two primary factors under investigation, medium of observation and previous experience of observer. During each of the nine experimental sessions observers were divided into three groups in the same manner as during training. Observers, in groups of three, were simultaneously situated either behind the one-way mirror, in the classroom, or in a separate room with a TV monitor of the classroom. Throughout the experimental sessions, three children, selected by the teacher as "typical," were observed by all nine observers for 12.5 min each. Therefore, during the course of the study, each of the children was observed 37.5 min *via* each medium each ses-

sion, and each observer provided recordings for 112.5 min for each child.

It was necessary for the three persons with their primary experience in each of the three media to contribute equally to recordings in each medium. It was further considered desirable to have each individual observe an equal number of times in each medium. The experimental design counterbalanced observers and groups during the nine sessions such that each observer participated in three different groups, and each group observed *via* each of the three media.

One announced reliability assessment was conducted each day to provide immediate feedback by a member of each three-person group who was identified after recordings for that 12.5-min period had been completed. A particular advantage of the present experimental design is that each observer served as assessor for each other observer in the study on only one occasion. This was intended to minimize between group differences due to "consensual drift" which would have been confounded with differences between media (O'Leary and Kent, 1973).

Reliability Computation

After all sessions were completed, the occurrence reliability of observers within each medium *and* between media was computed by dividing the number of agreements on the

occurrence of *each behavior* in *each recording interval* by the number of agreements plus disagreements. The recordings of each observer in each medium were compared an equal number of times with recordings of each other observer in each medium.

RESULTS

Recorded Frequencies

A 4-factor, repeated measures analysis of variance was employed in analysis of the behavioral frequencies, which are presented in Table 1. The between-subject factor was *training*, the exclusive experience of each observer prior to the study with one of the three media. The repeated measures (within-subject) factors were *media*, *target child*, and *replication*. The target children were the three individuals viewed throughout the study. Replication refers to the observation of each child in each medium on three occasions by each observer.

This analysis was performed separately for each of the nine categories of behavior and for a composite of all categories per 25 recording intervals. The 25 recording intervals were chosen as the basic unit for analyses since those intervals represented 12.5 min of observation time or the amount of time each child was observed by all nine observers. The main effect of media produced a significant effect on only one category, vocalization, $F(2, 12) = 9.10, p < .01$.

Table 1
Mean Behavioral Frequencies* Recorded in Each Medium

	Behavioral Categories									
	Out of Chair	Mod. Out of Chair	Touching	Vocalization	Playing	Orienting	Noise	Aggression	Time Off Task	Composite**
in vivo	3.62	3.80	0.44	11.44	7.81	5.30	7.74	0.14	4.19	44.75
mirror	3.52	4.15	0.83	7.52	7.44	4.88	6.28	0.14	4.09	38.74
television	3.10	4.49	0.49	6.73	7.81	5.15	7.37	0.05	4.20	39.33

*Frequencies are expressed as number of occurrences per 25, 20-second recording intervals.

**The composite score presented varies slightly from the sum across rows since daily composite scores were generated for each observer in each of the media and then a summary composite score was obtained.

Table 2
Reliability Coefficients within and between All Observational Circumstances

Reliability	Behavioral Categories									
	Out of Chair	Mod. Out of Chair	Touching	Vocalization	Playing	Orienting	Noise	Aggression	Time Off Task	Composite
<i>Within Medium</i>										
<i>in vivo</i>	.87	.55	.44	.80	.70	.66	.58	.33	.53	.71
mirror	.80	.62	.33	.60	.68	.72	.48	.75	.65	.69
television	.85	.71	.50	.63	.65	.65	.53	.50	.51	.65
<i>Between Media</i>										
<i>in vivo vs. mirror</i>	.63	.63	.25	.36	.68	.48	.49	.50	.37	.58
<i>in vivo vs. television</i>	.70	.41	.14	.36	.66	.59	.43	.50	.52	.55
<i>mirror vs. television</i>	.69	.37	.49	.40	.66	.47	.43	.00	.24	.50

Observers *in vivo* recorded an average of 11.44 vocalizations per 25 recording intervals as compared with 7.52 behind the mirror and 6.73 *via* television. A Duncan's multiple range test revealed that *in vivo* recordings of vocalizations were significantly different from those of mirror, $R_2(2, 12) = 3.65$, $p < .05$, and of television, $R_3(3, 12) = 4.44$, $p < .05$, but that recordings in the latter two media did not differ significantly. The main effect of training, as well as the interaction of training with media, target child, and replication, produced no significant effects on any of the nine behavioral categories or on the composite measure.

Reliability

Presentation of reliability data is limited to descriptive rather than inferential statistics which would be complicated by the non-normal distribution of reliability coefficients and complex interdependencies among these measures. Table 2 presents the average occurrence reliability for each category and the composite. The first three rows reflect reliability computed among observers recording simultaneously within the three media. The average reliability of the composite was .71 *in vivo*, .69 behind the mirror, and .65 *via* television. Reliabilities for individual categories in particular media

ranged from .33 to .87, with a median of .65.

Although analysis of variance of behavioral frequencies revealed lack of significant differences between media, reliabilities were computed to assess directly the similarity of recordings in the three media. Rows 4 through 6 of Table 2 present reliabilities among observers recording simultaneously in different media. The average reliability of the composite was .58 between *in vivo* and mirror, .55 between *in vivo* and television, and .50 between mirror and television. Reliabilities for individual categories ranged from .00 to .70, with a median of .49. The average reliability within media was higher for all nine categories than the average reliability between media.

Comparisons of composite reliabilities as calculated by four frequently used means, *viz.*, occurrence, occurrence and nonoccurrence, phi, and relative frequency, are presented in Table 3 (Hartmann, 1977; Kent and Foster, 1977, pp. 279-328).

DISCUSSION

The present investigation revealed little systematic distortion in frequency of recorded behavior as a function of different media or previous experience of the observer. The one excep-

Table 3

A comparison of composite reliabilities within and between all observational circumstances.

Reliability	Coefficient			
	Occur- rence	Nonoc- currence	Phi	Relative Fre- quency
<i>Within Media</i>				
<i>in vivo</i>	.71	.94	.80	.89
mirror	.69	.94	.77	.85
television	.65	.94	.74	.85
<i>Between Media</i>				
<i>in vivo vs.</i>				
mirror	.58	.89	.67	.88
<i>in vivo vs.</i>				
television	.55	.90	.66	.84
mirror vs.				
television	.50	.89	.59	.80

tion to this conclusion was vocalization. Differences in the media accounted for 9.68% of the variance of this category. As the category of noise was not similarly affected, this finding probably does not reflect a general liability for categories involving detection of sound. Rather, these data seem to replicate the difficulty encountered by Lovaas, Koegel, Simmons, and Long (1973) in obtaining sufficient fidelity for ratings of speech and facial expressions on videotapes.

Differences in the reliability of the three media were .06 or less on the composite of all nine categories. Comparisons of simultaneous recordings in different media, however, produced more substantial differences in reliability of the composite score, ranging from .13 to .21. Thus, the results indicate little difference in frequencies and reliabilities within a medium but clear loss of reliability between media. This suggests that recordings in different media may differentially reflect particular subcategories of behavior.

The use of an individual subject design in which observers rotate through various media conditions has certain distinct advantages and disadvantages that bear on the conclusions one

can draw from this study. Advantages of this method include use of relatively small numbers of observers, minimization of observer drift, and minimization of variability. Disadvantages of the design herein include possible awareness of the intermedia comparisons and possible interactive effects of alternately observing in varied modes, *viz.*, *in vivo*, *via* mirror, and *via* TV. For example, it is possible that observing *in vivo* prompts one to attend to behavior on a TV monitor to which one might not attend without prior *in vivo* experience. In brief, while these data would lead one to conclude that there are few if any critical differences in the data obtained through the three media, unequivocal statements about the comparability of the media as data sources await evaluations with independent groups.

It must be emphasized that the present data are based on a particular observation code and recording procedure applied to conduct-problem children in a classroom setting by experienced college student observers. Although similar data collection has occurred in a variety of clinical and methodological studies in the last decade, it should be clear that further evaluation of the effects of observation medium is needed in the context of other recording procedures, subjects, settings, and observers before generalizations can be drawn regarding the similarity of the three observation media. However, one can infer from the present data that the observation mirror and the closed circuit television may provide acceptable alternatives to *in vivo* observation.

REFERENCES

- Drabman, R. S., Spitalnik, R., and O'Leary, K. D. Teaching self-control to disruptive children. *Journal of Abnormal Psychology*, 1973, **82**, 10-16.
- Hartmann, D. P. Considerations in the choice of interobserver reliability estimates. *Journal of Applied Behavior Analysis*, 1977, **10**, 103-116.
- Kaufman, K. F. and O'Leary, K. D. Reward, cost, and self-evaluation procedures for disruptive adolescents in a psychiatric hospital school. *Journal of Applied Behavior Analysis*, 1972, **5**, 293-310.

- Kent, R. N., Diament, C., Dietz, A., and O'Leary, K. D. Expectation biases in observational evaluation of therapeutic change. *Journal of Consulting and Clinical Psychology*, 1974, **42**, 774-780.
- Kent, R. N. and Foster, S. L. Direct observational procedures: Methodological issues in naturalistic settings. In A. R. Ciminero, K. S. Calhoun, and H. E. Adams (Eds), *Handbook of behavioral assessment*. New York: Wiley, 1977.
- Lovaas, O. I., Koegel, R., Simmons, J. Q., and Long, J. S. Some generalization and follow-up measures on autistic children in behavior therapy. *Journal of Applied Behavior Analysis*, 1973, **6**, 131-166.
- Mercatoris, M. and Craighead, E. The effects of non-participant observation on teacher-pupil classroom behavior. *Journal of Educational Psychology*, 1974, **66**, 512-519.
- O'Leary, K. D., Drabman, R. S., and Kass, R. E. Maintenance of appropriate behavior in a token program. *Journal of Abnormal Child Psychology*, 1973, **1**, 127-138.
- O'Leary, K. D., Kaufman, K. F., Kass, R. E., and Drabman, R. S. The effects of loud and soft reprimands on the behavior of disruptive students. *Exceptional Children*, 1970, **37**, 145-155.
- O'Leary, K. D. and Kent, R. N. Behavior modification for social action: Research tactics and problems. In L. A. Hamerlynck, L. C. Handy, and E. J. Mash (Eds), *Behavior change: Methodology, concepts and practice*. Champaign, Illinois: Research Press, 1973.
- O'Leary, K. D., Kent, R. N., and Kanowitz, J. Shaping data collection congruent with experimental hypotheses. *Journal of Applied Behavior Analysis*, 1975, **8**, 43-51.
- O'Leary, S. G. and Schneider, M. J. Special class placement for conduct problem children. *Exceptional Children*, 1977, **44**, 24-30.

Received 11 January 1978.

(Final Acceptance 18 August 1979.)